Korpuszépítés ómagyar kódexekből

Simon Eszter, Sass Bálint, Mittelholcz Iván

MTA Nyelvtudományi Intézet
{eszter,sass.balint,mittelholcz}@nytud.hu

Kivonat Az annotált nyelvi erőforrások elérhetősége egyre fontosabb szerepet kap a nyelvészet több területén: a nyelvtechnológiai fejlesztéseken kívül az elméleti kutatásoknak is kiváló alapanyagot szolgáltatnak a korpuszok. A Magyar Generatív Történeti Szintaxis című projekt keretében felépítünk egy olyan korpuszt, amely tartalmazza az összes fennmaradt ómagyar szövegemléket. A cikkben a teljes korpuszépítési munkafolyamatot bemutatjuk – a szkenneléstől az online lekérdező felületig.

1. Bevezetés

Az annotált nyelvi erőforrások elérhetősége egyre fontosabb szerepet kap a nyelvészet több területén: a nyelvtechnológiai fejlesztéseken kívül az elméleti kutatásoknak is kiváló alapanyagot szolgáltatnak a korpuszok. A történeti korpuszok az adatok és a nyelvi jelenségek gazdag tárházát adják – de csak akkor, ha a releváns információ elektronikusan interpretálható és előhívható módon van tárolva bennük. A Magyar Generatív Történeti Szintaxis című projekt célja, hogy diakrón szintaktikai vizsgálatokat végezzen magyar nyelvű szövegeken, melyhez elsődleges fontosságú egy elektronikus nyelvtörténeti adatbázis létrehozása. A projekt időtartama alatt (2009-2013) felépítünk egy olyan korpuszt, amely tartalmazza az összes fennmaradt ómagyar kori (896–1526) szövegemléket, a középmagyar korból (1526–1772) pedig különféle szempontok szerinti arányos válogatást úgy, hogy minden nyelvjárás, műfaj, regiszter súlyának megfelelően képviselve legyen benne.

Napjainkban a korpuszépítési munkálatok során elsősorban már digitalizált szövegekből indulnak ki; de nem ez a helyzet a történeti dokumentumokkal. Az elektronikus formátumok (sőt az elektromosság) előtti korból származó szövegekből való korpuszépítés sokkal idő- és munkaigényesebb folyamat, és bizonyos esetekben más módszereket is igényel, mint a mai szövegek esetében. A tény, hogy az ómagyar kor több mint 6 évszázadot fog át, amelynek során nem volt egységes hangjelölési rendszer, vagyis az egyes szövegekben levő speciális karakterek halmaza különböző, tovább nehezíti a helyzetet. A helyesírás ezekben a századokban távolról sem volt egységes, ráadásul egy kódexet általában több kéz jegyez, ami még tovább növeli a heterogenitást a szövegekben. Ezek és más, később részletezett okok miatt a sztenderd előfeldolgozó lépések (tokenizálás, mondatra bontás, morfológiai elemzés és egyértelműsítés) nem végezhetők teljesen automatikusan, és nagyon sok kézi ellenőrzést igényelnek.

A cikkben a teljes korpuszépítési munkafolyamatot bemutatjuk – a szkenneléstől az online lekérdező felületig. A 2. fejezetben a korpusz anyagának összegyűjtését írjuk le, majd a 3. fejezetben bemutatjuk a korpusz felépítését, valamint az ezzel párhuzamos szövegfeldolgozási lépéseket. A 4. fejezet az online lekérdező felület leírását adja, végül a korpuszépítéssel kapcsolatos további feladatainkat tárgyaljuk.

2. A korpusz anyagának összegyűjtése

A reprezentativitás a korpuszok egyik lényegi tulajdonsága, kivéve abban az esetben, ha egy holt nyelvet vagy egy nagyon speciális nyelvi réteget vizsgálunk. Ez a helyzet az ómagyar korpusz esetében is, amely terveink szerint az összes ómagyar korból fennmaradt szövegemléket tartalmazni fogja. Szövegemlék alatt az összefüggő ómagyar mondatokat tartalmazó nyelvemlékeket értjük, az ún. szórványemlékekkel, amelyekben csak sporadikusan fordulnak elő magyar szavak vagy nevek, jelen projektben nincs lehetőségünk foglalkozni. Nem szerepelnek továbbá a korpuszban azok a szövegek sem, amelyeket még soha nem adtak ki nyomtatásban, vagyis a nyelvtörténeti átírási munkát is nekünk kellene elvégezni.

A fenti megszorításokat figyelembe véve a feldolgozandó ómagyar anyag 47 kódexet, 27 rövidebb szövegemléket és 244 misszilist (elküldött levelet) foglal magában, vagyis mindösszesen körülbelül 2 millió szövegszót. Ebből több mint 770 ezer már elérhető, kereshető állapotban van. A középmagyar kori szövegek kiválogatása még folyamatban van.

A korpuszépítés első lépése a valamilyen elektronikus szöveges formátumban már meglévő nyelvtörténeti anyagok összegyűjtése volt. A különböző forrásokból származó, változatos fontkészleteket használó, jellemzően Microsoft szövegszerkesztő eszközökkel előállított dokumentumokat egységes, UTF-8 kódolású, sztenderd Unicode-karaktereket tartalmazó sima szövegfájlokká alakítottuk. Egy másik forrásunk a Számítógépes Nyelvtörténeti Adattár volt, amelyben több ómagyar kódex ábécérendes adattára elérhető. A kódexfeldolgozási munkálatok még a hetvenes években kezdődtek a Debreceni Egyetemen Jakab László vezetésével. Az adattárban a kódex címszavai ábécérendbe rendezve szerepelnek. A hozzájuk tartozó betűhű szövegszavakat a lelőhely (lapszám, sorszám) megjelölésével közlik, mellettük számokkal rögzítették az adatra vonatkozó helyesírás-történeti, szótörténeti, hangtani, szófajtani, jelentéstani és alaktani tudnivalókat. Ez a fajta adatkódolási módszer még a hetvenes évekből maradt, mivel annak idején még lyukkártyán rögzítették az információkat. Ebből a táblázatos formából állítottuk vissza a kódexek eredeti betűhű szövegét, továbbá az egyes szövegszavakhoz tartozó morfológiai elemzést az általunk használt morfológiai elemző kimeneti formátumára átalakítva.

Az ómagyar szövegek nagy részének azonban nincsen elektronikusan elérhető szöveges változata, így ezeket a számítógép által olvasható és feldolgozható formára kell hoznunk. Ez a rövidebb szövegek esetében általában begépeléssel, a hosszabbak esetében szkenneléssel, optikai karakterfelismerő (OCR) program alkalmazásával és kézi ellenőrzéssel történik.

3. Az annotáció kidolgozása

Ahhoz, hogy a korpuszban a nyelvi jelenségek kereshetők legyenek, vagyis az adatbázis használható segédeszköze legyen az elméleti és nyelvtörténeti kutatásoknak, a releváns információknak elektronikusan interpretálható és előhívható módon kell tárolva lenniük. Ennek megvalósításához a sztenderd szövegfeldolgozó lépéseket (tokenizálás, mondatra bontás, morfológiai elemzés és egyértelműsítés) kell megtennünk, a történeti szövegek esetében azonban ezek nem problémamentesek. Bizonyos lépések automatizálhatók, de munkaigényesebb módszereket és több kézi ellenőrzést igényelnek, mint a mai nyelvet reprezentáló korpuszok esetében.

A korpusz felépítése, vagyis az egyes szövegszavakhoz tartozó annotációs szintek párhuzamosan alakulnak a szövegfeldolgozottsági szintekkel, melyeket az 1. táblázatban láthatunk. Ezek alapján hat annotációs szintet és öt feldolgozó lépést különíthetünk el, melyeket ebben a fejezetben ismertetünk részletesebben.

1. táblázat. Szövegfeldolgozottsági szintek.

(1) kiadott kódex szkennelve
→ OCR
(2) nyers OCR-kimenet
→ *kézi* javítás, kódolás
(3) betűhű elektronikus forma
→ *félautomatikus* normalizálás
(4) normalizált forma
→ *automatikus* morfológiai elemzés
(5) szótövesített és morfológiailag elemzett forma
→ *kézi* egyértelműsítés
(6) egyértelműsített korpusz

3.1. Szkennelés

Néhány kódex beszkennelt verziója megtalálható a Magyar Elektronikus Könyvtárban, sőt ezek egy része ún. „szendvics” PDF, vagyis a kép mögött megtalálható az OCR-ezett szöveg is. Ennek ellenére ezeket nem tudtuk használni: a képek felbontása nem elég jó az OCR-ezéshez, a mögöttes szöveg pedig nem esett át kézi ellenőrzésen, vagyis meglehetősen sok benne a hiba. Így minden kódexet, amit nem tudtunk szöveges formában megszerezni, minimum 300 dpi felbontásban be kellett szkennelnünk.

3.2. OCR

Az ómagyar kódexekben található nagyszámú különleges karakter kezelése miatt az OCR programmal szemben alapvető elvárásunk volt a taníthatóság. A

szóba jöhető nyílt forráskódú szoftverek (pl. Tesseract) tanítása túl időigényes lett volna, ezért végül az Abby FineReader mellett döntöttünk. Ez ugyan nem nyílt forráskódú, de meglehetősen könnyen tanítható, és elég jó minőségű kimenetet ad.

Az OCR program teljesítményét másokhoz hasonlóan (pl. [1]) nem karakterszinten, hanem szópontossággal (*word accuracy, WAcc*) mértük (az írásjelek felismerésétől eltekintettünk). Az előzetes elvárásoknak megfelelően az eredmények azt mutatják, hogy a pontosság nagyban függ a kódexekben alkalmazott helyesírástól. Kniezsa [2] az ómagyar kori kódexek kezeinek helyesírását három nagy típusba sorolja; a kiértékelésnél ezt a kategorizálást követtük. A mellékjel nélküli helyesírás a latinban nem szereplő magyar hangokat több betű kombinációjával írja le; a mellékjeles helyesírás egy rokonhang betűjének mellékjeles változatával jelöli ezeket; a harmadik típus pedig ezek keveréke. A kiértékeléshez három kódexet választottunk a három különböző típusból, továbbá összehasonlítási alapként egy rövidebb mai magyar szövegen is kiértékeltük a szoftver teljesítményét.

A legjobban a mellékjel nélküli helyesírással boldogult a program: ez nagyjából megegyezik a mai magyar szövegek felismerésében nyújtott pontossággal. A mellékjeles és keverék helyesírású kódexekben használt speciális karakterek nagy száma a tanítás ellenére is kb. 30%-kal rontotta a pontosságot.

2. táblázat. Az OCR szópontossága helyesírási típusok szerint.

kódex	helyesírás	tokenszám	felismert	WAcc (%)
Kulcsár	mellékjel nélküli	36.321	35.258	97,07
Müncheni	mellékjeles	74.657	50.790	68,03
Czech	keverék	11.478	7.910	68,91
–	mai magyar	5.121	5.068	98,97

3.3. A betűhű szöveg

A betűhű szöveg elkészítésekor nem a kódexek kézzel írott változatát, hanem az általunk használt átirat szerkesztőjének konvencióit követjük, vagyis nem törekszünk tökéletes paleográfiai pontosságra. A szabványosság előnyei miatt a teljes korpuszt sztenderd UTF-8 kódolású Unicode karakterekkel tároljuk és jelenítjük meg. Mindenképpen szükséges egy, az egész korpuszra kiterjedő szigorúan egységes formátum, ez teszi lehetővé, hogy a lekérdezéseket az egész anyagra vonatkoztathassuk. Ugyanakkor viszonylag nagy erőfeszítést kíván ennek az egységességnek a megvalósítása, mivel az egyes nyelvemlékek írásmódja, a bennük előforduló speciális ómagyar karakterek halmaza meglehetősen különbözik egymástól. A különféle ékezetes és többszörösen ékezetes karaktereket a Unicode megfelelően kezeli, de előfordulnak olyan régi magyar karakterek is, melyek a Unicode-ban nincsenek reprezentálva. Ezeket a karaktereket egy kiválasztott

Unicode karakterrel helyettesítjük, mégpedig úgy, hogy az adott helyettesítő karaktert kizárólag az adott hiányzó eredeti karakter helyett használjuk a korpuszban.

3.4. Normalizálás

A magyar írásosságot a latin nyelvű és vallásos tárgyú irodalom fordításának igénye hívta életre, de a latin ábécé magyarra alkalmazása számos problémát vetett fel. A legfőbb gond abból fakadt, hogy nyelvünk hangrendszerének több eleme a latinban ismeretlen, így ezek jelölésére új jeleket kellett bevezetni. Az ómagyar korban a helyesírás még egyáltalán nem volt egységesítve, sőt egy kódexet akár több kéz is jegyezhetett, ami további egyenetlenségeket okoz a szövegekben. A különböző helyesírási rendszerekben is ritka az egy hang–egy betű megfelelés (vagyis amikor egy hang jelölésére mindig ugyanaz a betű használatos, és az adott betűnek mindig egy hangértéke van), de egy alakulóban levő helyesírási rendszerben ilyenfajta következetesség még annyira sem várható el. Sőt inkább az a tipikus, hogy egy emléken belül is ingadozik egy-egy hang jelölésmódja (pl. *Vylag uilaga* [világ világa])), vagy kettős hangértéke van egy-egy betűnek (pl. *zerzete zerent* [szerzete szerint]). Tovább bonyolítja a helyzetet, hogy néhány betű egyaránt utalhat magánhangzóra és mássalhangzóra is (pl. az *u,v,w* több évszázadon át jelölhette az *u,ú,ü,ű,v* hangok bármelyikét).

Ezért szükség van egy ún. *normalizálás*i lépésre, amelynek során az eredeti betűhű szóalakokat mai magyar helyesírású szavakra alakítjuk át. A többféle, különböző nyelvtörténeti szakmai érvekkel alátámasztható lehetséges feldolgozási forgatókönyvek egyik gyakori közös átalakító lépése ez a fajta normalizálás (pl. [3]). A szövegfeldolgozásnak ez a lépése kritikus fontosságú, enélkül ugyanis a (félig) automatikus annotáció hatékonysága a következő lépésekben drámaian visszaesik [4].

Mivel a normalizálás nyelvtörténeti szakértelmet kívánó, rendkívül időigényes manuális munka, megpróbáltuk kiváltani gépi eljárással. Az általunk épített gépi normalizáló az ómagyar tokenekhez átírási lehetőségeket rendel, melyek közül a normalizálást végző nyelvész ki tudja választani a megfelelő kimenetet (részletesen lásd [5]).

A normalizálás során két alapelvet tartunk szem előtt. Egyrészt a ma nem létező összes szót, toldalékot, morfológiai konstrukciót megtartjuk, vagyis morfémát nem toldunk be, és nem hagyunk el. Másrészt viszont elhagyunk minden fonológiai és helyesírási esetlegességet, vagyis egységes, amennyire lehet, a mainak megfelelő helyesírásra törekszünk. Ez utóbbi azt is jelenti, hogy egy adott szót mindig ugyanúgy írunk le – ezt nevezzük az egységesség elvének.

A normalizálási lépés során történik meg a szöveg tokenekre és mondatokra való bontása is – mindkettő kézzel. Tokenizáláson jelen esetben azt értjük, amikor az ómagyar szövegben a szavakat a mai helyesírásnak megfelelően összevonjuk, illetve szétválasztjuk, természetesen a megfelelő módon jelölve a változtatásokat. Mivel ebben a korban a mai írásjelek nagy része még ismeretlen volt, továbbá amit használtak, azt se következetesen tették, a mai értelemben vett automatikus

mondatra bontás teljesen lehetetlen vállalkozásnak tűnik. Ezért ezt a szövegfeldolgozási lépést is manuálisan végezzük el.

3.5. Morfológiai elemzés és egyértelműsítés

A normalizált szövegváltozat képezi a morfológiai elemző bemenetét. Mivel a normalizálás során az ómagyar szöveget mai magyarra írjuk át, az ez utóbbira kifejlesztett automatikus morfológiai elemzőt viszonylag könnyen tudjuk alkalmazni a nyelvemlékek feldolgozására. Jelen projektben a *Humor* elemzőt használtuk [6]. Az egyik normalizálási alapelvünk, hogy minden morfológiai konstrukciót megtartunk, ezért természetesen ki kellett bővítenünk a lexikont és a szabályhalmazt bizonyos ma már nem létező, de az ómagyarban még használt nyelvi jelenségek leírásával. A morfológiai elemző kimenetének egyértelműsítését viszont – a gépi normalizáló kimenetének kezeléséhez hasonlóan – kézzel végezzük.

4. Korpuszlekérdező eszköz

A korpusszal párhuzamosan készül a hozzá tartozó korpuszlekérdező rendszer, amelynek segítségével a teljes ómagyar korpuszt kutathatjuk. A jó korpuszlekérdező eszközök lehetővé teszik azt, hogy kifinomult, nyelvészetileg releváns lekérdezéseket fogalmazzunk meg általuk. Az ilyen lekérdezések sok esetben különféle nyelvi szinteken megjelenő információra hivatkoznak. Hogy ez megvalósulhasson, adatbázisunk párhuzamosan tartalmazza az 1. táblázatban látható hat szövegfeldolgozottsági szintnek megfelelő nyelvi adatokat. Ezenfelül lehetővé tesszük a több szintre való egyidejű hivatkozást akár egy kérdésen belül is. Ha például az a kérdésünk, hogy milyen szavak szerepelnek egy igealak és egy igekötő között, akkor az elemzések szintjén (6) kell megfogalmazni a kérdést. Ha gyakorisági listát készítünk a korpusz egy részéből, akkor ezt megtehetjük például a szótövekből kiindulva, de rá lehet kérdezni közvetlenül az *nç* végű szavakra is, ekkor a (3) szinthez fordulunk.

A korpusztalálatok megjelenítése független a lekérdezéstől, abban az értelemben, hogy igény szerint bármilyen – akár a lekérdezésben nem is szereplő – szövegfeldolgozottsági szintet is megjeleníthetünk.

A korpusz anyaga vertikális fájlok formájában készül el. Ezek `.csv` formátumú táblázatok, melyek soronként egy szövegszót tartalmaznak, az egyes szövegfeldolgozottsági szintekhez tartozó információt pedig a megfelelő oszlopban, kiegészítve egy ,,Értelmezés" és egy ,,Megjegyzés" oszloppal. Ezt a formát XML-lé alakítjuk, így végezzük el a validációs lépéseket, melyek az adatbázis konzisztenciáját ellenőrzik. Egy következő átalakító lépés során alakul ki az alkalmas bemenet az *Emdros* [7] korpuszkezelő rendszer számára, melyre a lekérdezőfelület épül.

A lekérdező felület az 1. ábrán látható. A felület középső részén hivatkozhatunk az egyes szövegfeldolgozottsági szintekre. Az itt megadott adatokból az *OK* gomb megnyomására áll elő maga a lekérdezés a bal oldali szövegmezőben az Emdros lekérdezőnyelvén, ez szerkeszthető, és a *Mehet* gombbal futtatható.

1. ábra. A korpuszlekérdező felülete. A feltüntetett példában azokra a tokenekre keresünk, melyeknél a normalizált alak kezdete a *jonh* sztring.

2. ábra. Az 1. ábrán látható lekérdezés eredményének részlete: korpuszpozíciók, ahol a normalizált alak kezdete a *jonh* sztring.

Az 1. ábrán bemutatott lekérdezés eredménye a 2. ábrán látható. A találatok felett a lókuszjelölő található, mely a kódex azonosítójából, az oldalszámból és az adott szó egyedi azonosítójából áll. Az egyes találatokat táblázatos formában jelenítjük meg: a betűhű alak zölddel, a normalizált alak feketével, az értelmezés – az ómagyar *jonh* mai magyar megfelelője a *szív* szó – pedig kékkel.

Végül lássunk egy valódi ómagyar szintaxisra vonatkozó elméleti nyelvészeti kutatási kérdést, melynek megválaszolásához segítséget nyújthat a korpusz. A mai magyarban tagadás esetén az igekötő követi az igét (vö: *nem jön be*), az ómagyar viszont az igekötő + tagadószó + ige (vö: *be nem jön*) sorrendet használja legtöbbször. A szófajok sorozatára vonatkozó megfelelő lekérdezések a 3. ábrán láthatók. Ezt a jelenséget mutatja a Jókai-kódexből származó alábbi példamondat is: „Ver touaba **kÿ nem futott**” (Vér továbbá ki nem futott.).

Mai magyar szórend:

```
[W FOCUS w_6e ~ 'Mod']
[W FOCUS w_6e ~ 'V\.']
[W FOCUS w_6e ~ 'Vpfx']
```

Ómagyar szórend:

```
[W FOCUS w_6e ~ 'Vpfx']
[W FOCUS w_6e ~ 'Mod']
[W FOCUS w_6e ~ 'V\.']
```

3. ábra. A tagadott ige és igekötő sorrendi viszonyára vonatkozó lekérdezések. A `w_6e` jellemzővel a (6) szinten elérhető morfológiai elemzésre kérdezhetünk rá, a tagadószó kódja `Mod`, az ige kódja `V`, az igekötőjé pedig `Vpfx`.

A *Régi Magyar Konkordancia* nevet viselő lekérdezőfelület szabadon elérhető a `http://corpus.nytud.hu/rmk` címen.

5. További feladatok

Elsődleges feladatunk a teljes ómagyar anyag betűhű szöveges formában való előállítása és kereshetővé tétele. A normalizálást, valamint a morfológiai elemzést és egyértelműsítést csak a korpusz egy részén fogjuk végrehajtani.

Az ómagyar szövegek eleve adott heterogenitása mellett további problémákat okoz az is, hogy a különböző korokban kiadott nyomtatott kódexátiratok tipográfiai kényszerűségek miatt azonos karaktereket eltérően jelenítenek meg. Terveink között szerepel ezen esetlegességek kiküszöbölése, vagyis a különbözőképpen jelölt karakterek azonos sztenderd Unicode-karakterrel való lecserélése.

A középmagyar anyagok esetében már fontos szerepet játszik a reprezentativitás kérdése, ugyanis ebből a korból lényegesen több nyelvemlékünk származik, vagyis a teljes anyag feldolgozására ebben a projektben nem vállalkozhatunk.

A középmagyar szövegemlékek kiválogatásánál két fő szempontot tartunk szem előtt: csak a már szöveges formátumban elérhető dokumentumokkal foglalkozunk, és ezeket Dömötör [8] műfaji beosztását követve kategorizáljuk úgy, hogy minden regiszter megfelelően képviselve legyen a korpuszban.

Köszönetnyilvánítás

Az ómagyar korpusz építése a Magyar Generatív Történeti Szintaxis projekt keretében valósul meg. A projektet az OTKA NK 78074. számú pályázata támogatja. Köszönetet mondunk Novák Attilának, aki a morfológiai elemzést és a Jakab László-féle táblázatok átalakítását végzi.

Hivatkozások

1. Volk, M., Marek, T., Sennrich, R.: Reducing OCR Errors by Combining Two OCR Systems. In: Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010), Lisbon, Portugal, Faculty of Science, University of Lisbon (2010)
2. Kniezsa, I.: Helyesírásunk története a könyvnyomtatás koráig. Akadémiai Kiadó, Budapest (1952)
3. McEnery, T., Hardie, A.: Lancester Newsbooks Corpus. (2003)
4. Rayson, P., Archer, D., Baron, A., Culpeper, J., Smith, N.: Tagging the bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In: Proceedings of Corpus Linguistics, University of Birmingham (2007)
5. Oravecz, C., Sass, B., Simon, E.: Semi-automatic normalization of Old Hungarian codices. In: Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010), Lisbon, Portugal, Faculty of Science, University of Lisbon (2010)
6. Prószéky, G., Kis, B.: A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, Maryland, USA (1999) 261–268
7. Petersen, U.: Emdros – a text database engine for analyzed or annotated text. In: COLING 2004. (2004) 1190–1193
8. Dömötör, A.: Régi magyar nyelvemlékek. Akadémiai Kiadó, Budapest (2006)